Claims

1. A method of managing overload in a server system, having a service operating in response to input requests, and a server operation parameter related to the operation of said service, the method comprising the steps of :

a. monitoring successive values of the server operation parameter as a function of time,

b. from such values,

b1. evaluating a first condition, which involves whether the server operation parameter passes a first threshold value in a first direction, and

b2. evaluating a second condition, which involves whether the server operation parameter passes a second threshold value in a second direction, with the second direction being opposite to the first direction, and extending from the first threshold value to the second threshold value,

c. starting rejection of input requests, upon verification of a third condition, related to the verification of at least one of said first and second conditions, and

d. terminating rejection of input requests upon verification of a fourth condition, related to the verification of said second condition.

2. The method of claim 1, wherein the third condition of step c. comprises the fact the first condition has been verified, and the fourth condition of step d. comprises the fact the second condition has been verified.

3. The method of claim 1, wherein the third condition of step c. comprises the fact the second condition has not been verified during a grace period after the first condition has been verified, and the fourth condition of step d. comprises the fact the second condition has been verified after the third condition has been verified.

4. The method of claim 1, wherein step b1. is performed at a first rate, and step b2. is performed at a second rate, not lower than the first rate.

5. The method of claim 4, wherein step b2. is performed within a time period starting upon verifying the first condition at step b1., and terminating upon verifying the fourth condition at step d.

6. The method of claim 1, wherein said server operation parameter represents a quantity related to a memory usage in the server.

7. The method of claim 1, wherein said server operation parameter represents a quantity related to the server throughput and to the server latency.

8. The method of claim 7, wherein step a. comprises deriving the server operation parameter from a given combination of the server throughput with the server latency.

9. The method of claim 8, wherein step a. further comprises :

a1. maintaining a reference value of the server throughput and a reference value of the server latency, said reference values being updated upon verification of a fifth condition, comprising the fact that the current value of the server throughput does overlie its reference value, and

a2. deriving a reference value of said server operation parameter from a combination of the reference value of the server throughput with the reference value of the server latency, said combination being of the same nature as said given combination .

10. The method as claimed in claim 8, wherein said server operation parameter is derived from the ratio of the server throughput to the reference value of the server latency and said reference value of the server operation parameter is derived from the ratio of the reference value of the server throughput to the reference value of the server latency.

11. The method of claim 10, wherein the fifth condition further comprises the fact that the current value of the server latency does not overlie its reference value.

12. The method of claim 11, wherein the fifth condition further comprises the fact that the server requests queue length remains substantially constant.

13. The method of claim 9, wherein said first and second threshold values are derived from said reference value of the server operation parameter.

14. The method of claim 9, wherein steps a1. and a2. are performed at a third rate.

15. The method of claim 14, wherein the third rate is not lower than the first rate.

16. An overload manager device for use in a server system, having a service operating in response to input requests, and a server operation parameter related to the operation of said service, said device comprising :

- a monitoring function for evaluating successive values of the server operation parameter as a function of time,
- a first logic function capable of evaluating a first condition, which involves whether the server operation parameter passes a first threshold value in a first direction,
- a second logic function capable of evaluating a second condition, which involves whether the server operation parameter passes a second threshold value in a second direction, with the second direction being opposite to the first direction, and extending from the first threshold value to the second threshold value, and
- a request supervisor operable for :
    * starting rejection of the input requests, upon verification of a third condition, related to the verification of at least one of said first and second conditions, and
    * terminating rejection of the input requests upon verification of a fourth condition related to the verification of said second condition.

17. The device of claim 16, wherein the third condition comprises the fact the first condition has been verified, and the fourth condition comprises the fact the second condition has been verified.

18. The device of claim 16, wherein the third condition comprises the fact the second condition has not been verified during a grace period after the first condition has been

verified, and the fourth condition comprises the fact the second condition has been verified after the third condition has been verified.

19. The device of claim 16, wherein the first logic function is operable at a first rate, and the second logic function is operable at a second rate not lower than the first rate.

20. The device of claim 19, wherein the second logic function is operable within a time interval starting upon verifying the first condition, and terminating upon verifying the fourth condition.

21. The device of claim 16, wherein said server operation parameter represents a quantity related to a memory usage in the server.

22. The device of claim 16, wherein said server operation parameter represents a quantity related to the server throughput and to the server latency.

23. The device of claim 22, wherein the monitoring function is operable for deriving the server operation parameter from a given combination of the server throughput with the server latency.

24. The device of claim 23, further comprising a tracking function, said tracking function maintaining a reference value of the server throughput and a reference value of the server latency and being operable:
- for updating the reference values upon verification of a fifth condition, comprising the fact that the current value of the server throughput does overlie its reference value, and
- for deriving the first and second threshold values from a combination of the reference value of the server throughput with the reference value of the server latency, said combination being of the same type as said given combination.

25. The device as claimed in claim 23, wherein the server operation parameter is derived from the ratio of the server throughput to the reference value of the server latency and the

first and second threshold values are derived from the ratio of the reference value of the server throughput to the reference value of the server latency.

26. The device of claim 24, wherein the fifth condition further comprises the fact that the current value of the server latency does not overlie its reference value.

27. The device of claim 26, wherein the fifth condition further comprises the fact that the server requests queue length remains substantially constant.

28. The device as claimed in any of claim 24, wherein the tracking function is operable at a third rate.

29. The device of claim 28, wherein the third rate is not lower than the first rate.

30. The device of claim 16, comprising an overload manager object, having filter methods capable of implementing said request supervisor, a gauge monitor and further methods capable of implementing the monitoring function, the first logic function and the second logic function in cooperation with said gauge monitor.

31. The device of claim 30, wherein the overload manager object, the gauge monitor and the further methods are instanciated from at least one generic class.

32. The device of claim 31, wherein the overload manager object comprises at least one MBean.

33. The device of claim 21, comprising an overload manager object related to a memory usage in the server.

34. The device of claim 22, comprising an overload manager object related to the server throughput and to the server latency.

35. A portal server having an overload manager device as claimed in claim 21 having a service operating in response to input requests, and a server operation parameter related to the operation of said service, said device comprising :

- a monitoring function for evaluating successive values of the server operation parameter as a function of time,
- a first logic function capable of evaluating a first condition, which involves whether the server operation parameter passes a first threshold value in a first direction,
- a second logic function capable of evaluating a second condition, which involves whether the server operation parameter passes a second threshold value in a second direction, with the second direction being opposite to the first direction, and extending from the first threshold value to the second threshold value, and
- a request supervisor operable for :
    * starting rejection of the input requests, upon verification of a third condition, related to the verification of at least one of said first and second conditions, and
    * terminating rejection of the input requests upon verification of a fourth condition related to the verification of said second condition;

wherein said server operation parameter represents a quantity related to a memory usage in the server.

36. A computer readable medium including program instructions executable to implement a method of managing overload in a server system, having a service operating in response to input requests, and a server operation parameter related to the operation of said service, the method comprising the steps of :

a. monitoring successive values of the server operation parameter as a function of time,
b. from such values,
    b1. evaluating a first condition, which involves whether the server operation parameter passes a first threshold value in a first direction, and
    b2. evaluating a second condition, which involves whether the server operation parameter passes a second threshold value in a second direction, with the second direction being opposite to the first direction, and extending from the first threshold value to the second threshold value,

c. starting rejection of input requests, upon verification of a third condition, related to the verification of at least one of said first and second conditions, and

d. terminating rejection of input requests upon verification of a fourth condition, related to the verification of said second condition.